Critical reasoning on causal inference in genome-wide linkage and association studies

Li, Yang; Tesson, Bruno M.; Churchill, Gary A.; Jansen, Ritsert C.

Published in:
Trends in Genetics

DOI:
10.1016/j.tig.2010.09.002

Publication date:
2010

Opinion

Critical reasoning on causal inference in genome-wide linkage and association studies

Yang Li[1*], Bruno M. Tesson[1*], Gary A. Churchill[2] and Ritsert C. Jansen[1,3]

[1] Groningen Bioinformatics Centre, University of Groningen, The Netherlands
[2] The Jackson Laboratory, Bar Harbor, Maine, USA
[3] Department of Genetics, University Medical Centre Groningen, University of Groningen, The Netherlands

Genome-wide linkage and association studies of tens of thousands of clinical and molecular traits are currently underway, offering rich data for inferring causality between traits and genetic variation. However, the inference process is based on discovering subtle patterns in the correlation between traits and is therefore challenging and could create a flood of untrustworthy causal inferences. Here we introduce the concerns and show that they are already valid in simple scenarios of two traits linked to or associated with the same genomic region. We argue that more comprehensive analysis and Bayesian reasoning are needed and that these can overcome some of the pitfalls, although not in every conceivable case. We conclude that causal inference methods can still be of use in the iterative process of mathematical modeling and biological validation.

Causal inference from genetic data

Understanding how genes, proteins, metabolites and phenotypes connect in networks is a key objective in biology. Genes are transcribed and translated into proteins that can act as enzymes to convert precursor metabolites into product metabolites. These relationships are often depicted informally using graphs with arrows pointing in the assumed direction of causality, for example, from genes to proteins to metabolites to classical phenotypes. These diagrams reflect our assumptions about causality in biological systems and in many cases have been painstakingly validated in controlled experimental settings. Today, more than ever before, we are faced with large-scale 'post-genomics' data that have the potential to reveal a multitude of as yet unknown but potentially causal relationships.

Methods for causal inference have been introduced as early as the 1920 s [1] and have been further developed and applied since then in genetic epidemiology and other fields [2–4]. Causal inference is a formal statistical procedure that aims to establish predictive models. For example, if a reduction in the level of a crucial metabolite is the cause of a disease, then an intervention that increases the metabolite level should alleviate the disease. By contrast, if the reduced metabolite is a consequence of the disease, then

Glossary

Allele frequencies: at a given polymorphic locus, different alleles can differ in prevalence within the population studied. In GWLS using a cross originating from two inbred founders the QTL has two alleles at equal frequencies in the population under study. By contrast, due to a combination of random segregation, drift and selection, allele frequencies in GWAS can be markedly different from equal. Imbalanced allele frequencies are less optimal for QTL detection.

Causal anchor: causal relationships that are provided by knowledge external to the data. Because meiotic recombination is a random process that pre-dates the establishment of phenotypes, correlation between DNA variation (QTL) and a trait implies causation of the DNA variation on the trait variation in experimental populations: QTL can therefore be used as causal anchors. The assumption should be carefully evaluated in natural populations, which can have hidden structure, or in case–control studies where sampling could indirectly alter allelic associations.

Causal inference: a process of determining whether variation observed in a trait is a cause or a consequence of variation observed in another trait. Here we adopt the definition used in [3] that causality is defined by the effects of intervention in a system. If X is a cause of Y, then we can predict that an intervention that alters the level of X will result in a change in Y.

Correlation: a statistical measure of how much two variables change together. Correlation best captures linear relationships between variables (on the original scale or after transformation).

Genome-wide association studies (GWAS): an experiment in which the genomes of unrelated individuals are screened for genetic markers [typically millions of single nucleotide polymorphisms (SNPs)] at which allelic variation correlates with variation in studied traits.

Genome-wide linkage studies (GWLS): an experiment in which the genomes of related individuals are screened for genetic markers (typically a few hundreds or thousands of SNPs) at which allelic variation correlates with variation in studied traits. Examples of GWLS include experimental crosses such as recombinant inbred panels, intercrosses and backcrosses.

Prior (or prior probability): reflects the initial belief in a given proposition (such as 'trait T1 is causal for trait T2') before observing the data. The application of Bayes' rule combines the evidence provided by observed data with the prior to provide a measure of evidence of the proposition that accounts for previous experience or external knowledge.

Quantitative trait locus (QTL): a genomic region is said to be a QTL for a trait if allelic variation in this region correlates with trait variation. QTL can be mapped through GWAS or GWLS.

- **Distant eQTL:** a distant (or *trans*) eQTL is an eQTL which is located far from the gene it controls (for example on a different chromosome).

Corresponding author: Jansen, R.C. (r.c.jansen@rug.nl)
* Equal contribution.

- **Expression QTL (eQTL):** a region in the genome at which allelic variation correlates with the mRNA expression-level variation of a particular gene.
- **Local eQTL:** a local (or *cis*) eQTL is an eQTL which is located near to the gene it controls in the genome. Often a local eQTL will be caused by allelic variation in the regulatory region of the gene or within the gene itself.
- **Metabolite QTL (mQTL):** a region in the genome at which allelic variation correlates with the abundance variation of a particular metabolite.
- **Protein QTL (pQTL):** is a region in the genome at which allelic variation correlates with the abundance variation of a particular protein. As with eQTL, pQTL can be local or distant according to the genomic position of the gene encoding the protein relative to the QTL.
- **QTL confidence interval:** QTL mapping identifies regions of the genome in which allelic variation is linked or associated with a specific trait. The sample size, the density of available genotyped markers and the extent of recombination in the QTL region within the studied population are among the factors that influence the size of the confidence interval. Confidence intervals can extend from only a few kb to several Mb, complicating the identification of the actual polymorphism behind the QTL.
- **QTL–trait–trait triads:** a set constituted by a QTL and two traits mapping to that QTL. Because a QTL can affect directly a trait, or indirectly through another intermediary trait, multiple causal scenarios can explain this triad as illustrated in particular by the blue models in Figure 1. This article discusses our ability to discriminate between those different scenarios.

Regression: a statistical procedure which evaluates the dependence between a variable (e.g. a trait) and one or multiple other variables (e.g. another trait, or QTL genotypes).
Residuals: in a regression, residuals are the differences between the observed values and the values fitted by the regression.
Variance: a statistical parameter that quantifies the spread in the distribution of a variable. For phenotypic traits variance originates from both genetic and non-genetic sources and we can estimate the proportion of trait variance that is contributed by a given QTL.

intervention will not have the desired effect. Causal reasoning is thus crucial to the process of target discovery in pharmaceutical research.

Recent genome-wide linkage studies (GWLS) on model organisms [5–7] and genome-wide association studies (GWAS) on humans [8] have successfully connected molecular and classical traits into networks with arrows indicating inferred causal relationships [9–17]. Causality cannot be established from data alone. Some assumptions about the causal relationships among the variables being modeled are needed. Once these are established, causal inference can be propagated to additional variables. In GWLS and GWAS settings it is typical to assume that genomic variation (quantitative trait locus/loci, QTL; Glossary) acts as a causal anchor from which all arrows are directed outward. Although this assumption seems quite natural, caution is warranted when the sample is not random, as in case–control studies.

There are many possible causal networks even in a simple system consisting of a genomic locus (QTL) and two traits, T1 and T2 (Figure 1). Causal inference in GWLS and GWAS involves, in its simplest form, the identification of pairs of traits with a common QTL (QTL–trait–trait triads) and determining whether the QTL directly affects each of two traits (independent), or if the QTL affects only one trait which in turn affects the other trait (causal or reactive). If none of these situations apply we assume that the causation is more complex (undecided).

Biological variation in the two traits beyond that induced by common QTL is the key for distinguishing between the independent and causal scenarios. If there is a causal link, the biological and QTL variation from T1 will propagate to T2. If the variation propagates in an approximately linear fashion we can, with simple linear regression (Box 1),

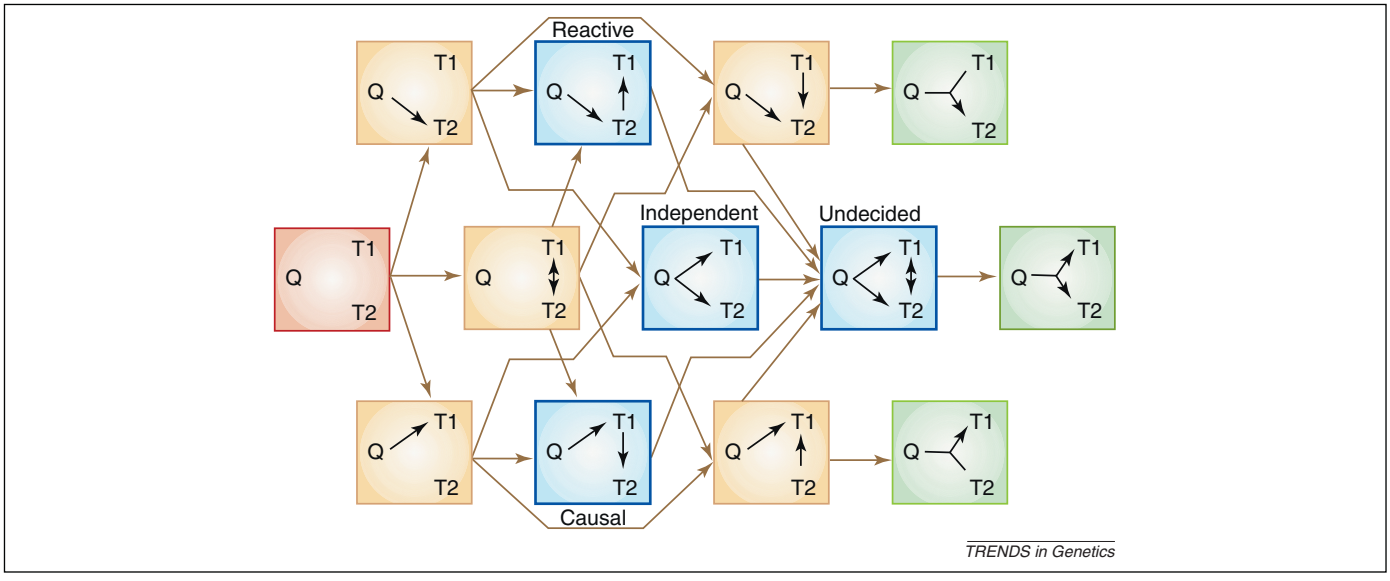

Figure 1. Triad models. Many different causal relationships are possible within a triad of two traits (T1 and T2) and a QTL (Q). The simplest case (red box) to the left shows no causality, in which case the QTL and the two traits do not influence each other. In the next set of models (yellow), at least one trait is not associated with the QTL. All these models are excluded from consideration based on the assumption that the QTL mapping step has correctly inferred the QTL–trait associations. The models that remain to be discriminated are highlighted in blue and green: the procedure to decide in favor of one of the blue causal topologies is outlined in the text. The three models furthest to the right (green) are extensions of the causal model that include additional interaction terms, for example the QTL could modulate the causal effect of T1 on T2. Equivalently, these models could be seen as relaxing the assumption of equal covariance across genotype classes. An extreme scenario is the Simpson's paradox model in which the traits show opposite correlations for different genotypes at the QTL. Such complexities are usually not considered, but could form an important part of actual biological networks. The brown arrows indicate which of the models are nested and can thus be compared directly by statistical testing.

Box 1. Causal inference with triads

(A) Decision procedure

The triad analysis is a statistical decision procedure consisting of the following steps:

Step 1: establish that two traits are linked to the same locus. This rules out the red and yellow models (Figure 1). We are ignoring the green models. We are thus now reduced to the four blue models (independent, causal, reactive, undecided).

Step 2: regress T2 on T1 and T1 on T2 to obtain residuals of each trait adjusted for the other. Denote residuals by R2 and R1, respectively.

Step 3: compute a bivariate *t*-test for association between the residuals (R1 and R2) and the QTL. Note that R2 is 100% adjusted for both QTL effect under the causal model only (zero expected value; Table I). We note that in other implementations of triad analysis one would compute univariate *t*-tests of R1 against QTL and R2 against QTL. This ignores the correlation between these two tests and we have amended it here.

Step 4: choose a model based on outcomes of the bivariate *t*-tests using a *P* value of for example 10%: independent if (yes, yes), causal if (yes, no), reactive if (no, yes). If none of these apply we default to the 'undecided' case.

(B) Properties of procedure

We describe two statistical measures and derive implications for population size:

Sensitivity: the sensitivity of the method is the probability of correctly detecting a true causal relationship. This probability is obtained from the non-central bivariate *t*-distribution (QTL effect of residuals determine the non-centrality; Table I).

Positive predictive value: the probability of a declared causal connection being true. We incorporate prior knowledge (Box 2 and Glossary): P1 is the product of the prior probability of a link to be causal times the probability to correctly identify a causal link as such; P2 is the product of the prior probability of a link to be independent times the probability to incorrectly identify an independent link as causal. The positive predictive value is then P1 / (P1 + P2).

Required population size: the above process is repeated for all combinations of QTL variance in the two traits, and for sample sizes ranging from 200 to 51200. The minimum sample size to achieve both 50% sensitivity and 90% positive predictive value is plotted (Figure 2).

Table I. Equations for regression parameters in the basic independent and causal model (first scenario in the main text)[a,b]

		Independent model	Causal model
		T1 = QTL + e1	T1 = QTL + e1
		T2 = QTL + e2	T2 = T1 + e2
Regress T1 on T2	Slope	$1 - v_2/v_{t2}$	$1 - v_2/v_{t2}$
Regress residual R1 on QTL	QTL effect	$2v_2/v_{t2}$	$2v_2/v_{t2}$
	Variance[c]	$v_1 + v_2(v_2/v_{t2} - 1)^2$	$v_2(v_2/v_{t2} - 1)^2 + v_1(v_2/v_{t2})^2$
Regress T2 on T1	Slope	$1 - v_1/v_{t1}$	1
Regress residual R2 on QTL	QTL effect	$2v_1/v_{t1}$	0
	Variance[c]	$v_2 + v_1(v_1/v_{t1} - 1)^2$	v_2
Covariation of QTL effects	Covariance[c]	$v_1(v_1/v_{t1} - 1) + v_2(v_2/v_{t2} - 1)$	$v_2(v_2/v_{t2} - 1)$

[a]T1 and T2 have mean zero and equal QTL effect; this can always be achieved by subtracting the means and rescaling.

[b]Here, e1 and e2 represent variance in the biological process, not measurement errors; v_1 and v_2 denote the variances of e1 and e2; and v_{t1} and v_{t2} denote the total variance which is sum of the QTL and the biological variances. The ratio v_1/v_{t1} is the proportion of total variance that is not explained by the QTL.

[c]Multiply by $1/n_A + 1/n_B$ in case of two genotypes where n_A (n_B) is the number of samples with genotype A (B); multiply by $4n / [n(n_A + n_B) - (n_A - n_B)^2]$ in case of three genotypes where $n = n_A + n_H + n_B$ is the total number of samples. Note that $4n / [n(n_A + n_B) - (n_A - n_B)^2] = 1/n_A + 1/n_B$ if $n_H = 0$.

subtract the biological and QTL variation in T1 from T2, and we are left with the additional or 'residual' variation in T2 that is unrelated to the QTL. If we attempt the reciprocal analysis, the additional variation in T2 could make the linear regression fail to subtract all of the QTL variation from T1. As a result the residual variation in T1 will still relate to the QTL. This reasoning suggests a simple approach for distinguishing between the independent and causal models on the basis of the outcome of two reciprocal statistical tests: does the residual variation in T1 still relate to the QTL, and does the residual variation in T2 still relate to the QTL. Traits are declared independent (yes, yes), causal (yes, no), reactive (no, yes), or more complex (no, no) in which case no decision is made (Box 1 for statistical details). Although the apparent simplicity of this approach is seductive, here we highlight some possible pitfalls illustrated by three simple but realistic scenarios, and discuss avenues to restoring the potential of causal inference.

Concerns about causal inference

It is compelling to explore how this causal inference method for QTL–trait–trait triads performs, particularly in GWAS where the majority of QTL identified explain much less than 5% of the total variance [18]. The method will declare particular triads to be independent and others to be causal, but such inferences are not without error. Of all triads that are truly causal, what proportion can be correctly identified as such? This proportion is referred to in statistics as the 'sensitivity' of the method. It is good for a method to be sensitive, but not sufficient to make it of practical use. Triads with truly independent traits can in some cases be incorrectly identified as causal by the method. As a consequence, the potential number of false causal links arising from, say, 80% of independent trait–trait pairs can overwhelm the number of true causal links arising from the 20% of causal trait–trait pairs. The proportion of true causal links among those identified as causal is referred to in statistics as the 'positive predictive value'. A good method combines a high positive predictive value, say 90%, with an acceptable sensitivity, say 10% or higher (Box 1 for statistical details). A QTL is a genomic region that can contain multiple candidate genes and polymorphisms. Without prior knowledge that two traits sharing a common QTL are biologically or biochemically related, they are more likely to be regulated by different genes or polymorphisms within the QTL region. In which case we would say the traits are independent and that their apparent relationship is explained by linkage disequilibrium and not by a shared biological pathway. Different types of prior knowledge about the (unknown) number of true causal and true independent relationships can be incorporated into the causal inference (Box 2).

Box 2. Bayesian Reasoning

Bayes rule [33] is a probability property that allows one to combine evidence from data with existing knowledge and expertise through the inclusion of priors in an inference process. The definition of the prior in a causal inference on a QTL–trait–trait triad is the result of a partly subjective process that can be guided by the following considerations:

- **QTL confidence interval size.** The larger the confidence intervals of the QTL are, the more likely it is that distinct polymorphisms control the traits. Linkage disequilibrium is pervasive in GWLS, leading to large confidence intervals.
- **SNP density in the QTL region within the population**. The more polymorphic the QTL region is, the more likely it is that the traits are actually controlled by distinct polymorphisms. In GWAS, populations are heterogeneous leading to a lot of allelic diversity along the genome.
- **Gene density within the confidence interval**. Polymorphisms that lie within gene coding regions are more likely to propagate variation at phenotypic level than polymorphisms in non-coding regions. The fewer the number of genes within the QTL confidence interval, the more likely that the two traits are affected by the same polymorphism.
- **Local or distant eQTL.** If a gene expression trait is locally regulated by an eQTL and the other trait is distantly regulated by the eQTL, then the gene with the local eQTL is more likely to be causal for the other trait than the other way around [14].
- **Additional shared QTL.** The sharing of multiple additional QTL between the two traits could be taken as additional evidence that they are connected in the network [31]. It is more likely that these QTL affect the traits through the same polymorphisms than it is that locations of multiple distinct polymorphisms coincide by chance.
- **QTL hotspot.** Regions of the genome, known as QTL hotspots, have been reported that harbor QTL for large numbers of traits. These could be the result of a single major polymorphism or of many polymorphisms in linkage disequilibrium and each affecting different traits independently. Further investigation and experience in understanding this phenomenon is needed to determine which is more likely.
- **Independent biological knowledge.** Biological knowledge about the two traits (for example if the two genes belong to a same KEGG pathway) can be used as *a priori* evidence that the traits are related.

We present three different scenarios to illustrate the properties of the method. In the first scenario T1 is causal for T2, all QTL and biological variation in T1 is propagated to T2 and, on top of this variation, T2 shows additional variation. This additional variation can originate from an independent perturbation, such as another QTL affecting T2 but not T1, or from an environmental perturbation affecting T2 but not T1. The correlation between T1 and T2 results fully from the causal relationship between the two traits. Exact analytical equations can be used to compute the population size required to attain the desired levels of sensitivity and positive predictive value (Box 1). This requires specifying the size of the QTL effect, the frequency in the population of the major QTL allele, and the prior belief that the triad is causal rather than independent. A population size of approximately 200–6000 (GWLS) to 800–25000 (GWAS) provides 50% sensitivity and 90% positive predictive value for causal inference, with QTL explaining from 30% down to 0.5% of total variance (Figure 2, with parameters as specified in the legend). Lowering the sensitivity to 10% would reduce the required population size, but this effect is visible only in the area close to the diagonal (Figure 2). In this area traits are too tightly correlated and there is little additional variation in T2, making it difficult to infer the correct causal direction, in other words sensitivity is low.

In the second scenario one or more shared hidden factors cause additional correlation between the traits. One can think of undetected QTL with pleiotropic effects on the traits, such as structural chromosomal variation leading to coexpression of genes in a particular region, physiological variation related to daily circadian rhythms, or environmental variation due to features of the experimental implementation. In a causal model the effect of the hidden factor acts on T2 in two ways: indirectly through T1, but also directly. For increasing values of hidden factor correlation (keeping QTL and total variance constant) the linear regression will tend to subtract the effect of the hidden

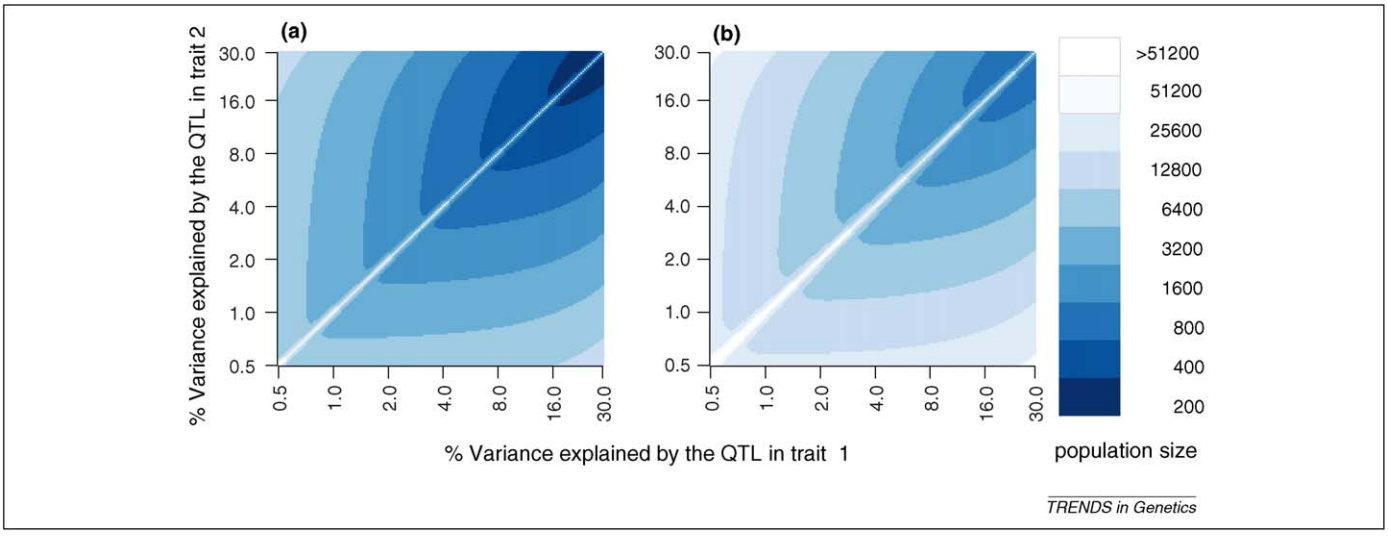

Figure 2. Population sizes required for reliable causal inference. Here we show the required population sizes in **(a)** GWLS, and **(b)** GWAS. Each color represents a different population size; the scale is shown in the right panel. These numbers have been calculated from the equations in Box 1 by using a 10% significance threshold for the *t*-tests, 90% positive predictive value and 50% sensitivity. We assume that there is only biologically variation and no measurement error. The x (or y) axis indicates the percentage of variance explained by a QTL in trait T1 or T2, respectively, on a logarithmic scale ranging from 0.5% to 30%. Allele frequencies of the biallelic QTL are set equal in GWLS, and at 10% and 90% in GWAS. Furthermore we use Bayesian reasoning (Box 2): we assume *a priori* that only 1% (20%) of the QTL–trait–trait connections is truly causal in GWLS (GWAS).

factor and not that of the QTL. As a consequence the causal links can appear to be independent (yes, yes); increasing sample size will not help to attain the desired levels of sensitivity and positive predictive value**.** In an independent model, the effect of the hidden factor acts on T1 and T2 directly, and not indirectly. As with the causal model, for increasing values of hidden-factor correlation (keeping QTL and total variance constant), the linear regression will typically tend to subtract the effect of the hidden factor and not that of the QTL. However, in the special case of equal slopes for hidden-factor and QTL, the linear regression will be able to subtract hidden factor and QTL effects. A truly independent model then tends to change from correct identification (yes, yes) via either causal (yes, no) or reactive (no, yes) to undecided (no, no). Increasing sample size will help only when slopes are still slightly different, but not if they are equal. Note that equal slopes cannot occur in the causal model, because the hidden factor acts directly and indirectly on T2. Sample size shown in Figure 2 is still approximately adequate if the hidden-factor variance is small, in other words equals at most the QTL variance.

In the third scenario, measurement error comes into play, which is realistic for most technologies for scoring molecular and classical traits. Note that the use of surrogate variables, such as RNA expression as a proxy for the causal protein levels, can also introduce a kind of measurement error. Measurement variation is never 'biologically' propagated from one trait to another trait, but it will change (reduce or increase) the correlation between the two traits, and thus the causal inference will be affected. Correlated measurement errors are analogous to the hidden-factor scenario described above with one exception. The special case of equal slopes for hidden factor and QTL can now occur also in the causal model: slopes for correlated measurement error and QTL can be equal. In this case, a true causal model can change from correct identification (yes, no) to undecided (no, no). Independent measurement errors will cause the linear regression to fail to subtract the QTL variation in both reciprocal analyses; therefore the causal model will tend to appear to be independent (yes, yes) if the measurement variance increases. However, an actual causal link from one trait measured with large measurement error to a downstream trait measured with small measurement error can be reported as reactive [13]. Again, increasing sample size will not be helpful to attain the desired levels of sensitivity and positive predictive value.

Restoring the potential of causal inference

We have explored causal inference in the simple context of QTL–trait–trait triads using a statistical decision procedure (Box 1) to potentially reject the undecided model in favor of one of the nested causal, reactive and independent models. This procedure is similar to other implementations of triad analysis [5,7,9] which, although not identical, lead to comparable results [11]. Other computational methods for causal inference such as structural equation modeling [19,20] or Bayesian network analysis [21] can operate on larger numbers of traits and QTL. These methods also rely on the correlation structure in the data and will therefore suffer from some of the same problems as triad analysis: they require large population sizes, and can be confounded by hidden factors or measurement noise. This calls for several recommendations to restore the potential of causal inference.

Our first recommendation is to use Bayesian reasoning in the causal inference procedure. Prior belief or knowledge about the number of true causal and true independent links that might be expected in a typical QTL, depending on the study design, should be considered to safeguard against high false-positive rates (low positive predictive values). In studies that involve mapping gene expression (eQTL), protein (pQTL) or metabolite (mQTL) traits, information about colocalization of QTL and genes that are functionally linked to the trait provides information about the likelihood of causal links. Lastly, biological annotations such as Gene Ontology [22] or Kyoto Encyclopedia of Genes and Genomes (KEGG) [23] pathways should also be considered when weighing evidence for causal links. The use of more informative priors (Box 2) provides better prioritizing and filtering of the large numbers of possible triads, and could reduce the population sizes required for reliable causal inference to more realistic numbers.

Our second recommendation is to identify and eliminate or account for experimental factors that can induce spurious correlation. It is not usually possible to measure all relevant factors, but even some of the most obvious factors such as age or sex of study subjects are often not taken into account. Any variation in diet, time since last feeding or time of sample collection, the size of plant seeds or the size of litter, temperature and light cycles, location in the greenhouse or field, can have profound effects. Such factors can be easily included in the model, but only when they are recorded [24,25]. Although it might not be necessary in inbred line crosses studies, it is crucial to consider the impact of population structure in almost every other setting where genetic variation is present. Methods are available to estimate kinship and the corresponding structure of the correlation. Combining these methods with causal inference can minimize the effects of spurious genetic correlation [26]. The effects of hidden factors affecting larger numbers of traits can be detected and corrected for by dimension-reduction methods [26–30]. Causal inference can then be applied to the residual data. However, these multivariate analysis methods also have the potential to remove from the data signals that are relevant for causal inference, and their application should be considered carefully.

Our third and final recommendation is to consider a richer set of possible models than the four blue models in Figure 1. For example, fitting a model such as the top-right yellow model in Figure 1 could provide a powerful case for a causal signal in the data [17,19,20]. The green models in Figure 1 with more complex correlation structure can also be informative and have been explored [17]. If two traits have multiple QTL in common, then this can be taken as additional evidence that the two traits are connected in the network [31]. This allows for the possibility to generalize the triad analysis to a multiple QTL–trait–trait analysis. A test of the effects of all QTL that propagate from one trait to another can be obtained by modifying step 3 in the decision procedure (Box 1) to assess the combined effect [32].

Concluding remarks

Many in the scientific community share a healthy skepticism of causal inference and, as we have shown, for good reasons. Nevertheless we conclude that causal inference in linkage or association analysis could soon become a feasible strategy given the rapidly growing prior knowledge of biological networks, the increasing population sizes, the advent of cheaper and more accurate measurement techniques, and the possibility of coupling causal inference methods with Bayesian reasoning. Further development of methods that consider the simultaneous effects of multiple traits and multiple QTL is needed, as well as the development of techniques that address the effects of experimental factors, study design and population structure. Reasonable caution is still warranted, and statistical methods of causal inference should be viewed as a necessary step in an era of high-throughput data generation and discovery.

Acknowledgements

This work was funded by 7th Framework Programme of the European Commission under the Research Project PANACEA, Contract No. 222936 to Y.L., and by the BioRange programme from the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK (Investments in Knowledge Infrastructure Directive) grant through the Netherlands Genomics Initiative (NGI) to B.M.T.

References

1 Wright, S. (1921) Correlation and causation. *J. Agric. Res.* 20, 557–585
2 Duffy, D.L. and Martin, N.G. (1994) Inferring the direction of causation in cross-sectional twin data: theoretical and empirical considerations. *Genet. Epidemiol.* 11, 483–502
3 Pearl, J. (2000) *Causality: Models, Reasoning, and Inference*, Cambridge University Press
4 Spirtes, P. *et al.* (1993) *Causation, Prediction, and Search*, Springer-Verlag
5 Chen, Y. *et al.* (2008) Variations in DNA elucidate molecular networks that cause disease. *Nature* 452, 429–435
6 Zhu, J. *et al.* (2008) Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat. Genet.* 40, 854–861
7 Schadt, E.E. *et al.* (2005) An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* 37, 710–717
8 Emilsson, V. *et al.* (2008) Genetics of gene expression and its effect on disease. *Nature* 452, 423–428
9 Chen, L.S. *et al.* (2007) Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biol.* 8, R219
10 Aten, J.E. *et al.* (2008) Using genetic markers to orient the edges in quantitative trait networks: the NEO software. *BMC Syst. Biol.* 2, 34
11 Millstein, J. *et al.* (2009) Disentangling molecular relationships with a causal inference test. *BMC Genet.* 10, 23
12 Chaibub Neto, E. *et al.* (2008) Inferring causal phenotype networks from segregating populations. *Genetics* 179, 1089–1100
13 Rockman, M.V. (2008) Reverse engineering the genotype–phenotype map with natural genetic variation. *Nature* 456, 738–744
14 Zhu, J. *et al.* (2004) An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet. Genome Res.* 105, 363–374
15 Bing, N. and Hoeschele, I. (2005) Genetical genomics analysis of a yeast segregant population for transcription network inference. *Genetics* 170, 533–542
16 Li, H. *et al.* (2005) Inferring gene transcriptional modulatory relations: a genetical genomics approach. *Hum. Mol. Genet.* 14, 1119–1125
17 Kulp, D.C. and Jagalur, M. (2006) Causal inference of regulator–target pairs by gene mapping of expression phenotypes. *BMC Genomics* 7, 125
18 Visscher, P.M. *et al.* (2008) Heritability in the genomics era – concepts and misconceptions. *Nat. Rev. Genet.* 9, 255–266
19 Li, R. *et al.* (2006) Structural model analysis of multiple quantitative traits. *PLoS Genet.* 2, e114
20 Liu, B. *et al.* (2008) Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics* 178, 1763–1776
21 Zhu, J. *et al.* (2007) Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLoS Comput. Biol.* 3, e69
22 Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29
23 Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic. Acids Res.* 28, 27–30
24 Li, Y. *et al.* (2008) Generalizing genetical genomics: getting added value from environmental perturbation. *Trends Genet.* 24, 518–524
25 Akey, J.M. *et al.* (2007) On the design and analysis of gene expression studies in human populations. *Nat. Genet.* 39, 807–808
26 Kang, H.M. *et al.* (2008) Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics* 180, 1909–1925
27 Dubois, P.C. *et al.* (2010) Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* 42, 295–302
28 Fehrmann, R.S. *et al.* (2008) A new perspective on transcriptional system regulation (TSR): towards TSR profiling. *PLoS One* 3, e1656
29 Leek, J.T. and Storey, J.D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* 3, 1724–1735
30 Stegle, O. *et al.* (2010) A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.* 6, e1000770
31 Jansen, R.C. and Nap, J.P. (2001) Genetical genomics: the added value from segregation. *Trends Genet.* 17, 388–391
32 Sargon, J.D. (1958) The estimation of economic relationships using instrumental variables. *Econometrica* 26, 393–415
33 Stephens, M. and Balding, D.J. (2009) Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* 10, 681–690